CSCI-B 565 DATA MINING
Homework 1
Computer Science Core
Fall
Indiana University,
Bloomington, IN

Shridivya Sharma
shrishar@umail.iu.edu

9/23/2016

All the work here in is Solely mine

k-means

1. **Answer 1**
   It's not necessary that k-means always converge.We try to obtain local minima(by choosing mean as centroid which ensure to obtain least SSE) in k- Means which doesn't always guarantee a global minima . It fails to converge in following situations

   (a) Non-spherical data where mean value cannot converge towards cluster center.i.e data is non-convex.

   (b) Data set having clusters of different size and density .

   In this kind of situation, We need a parameter which represents **maximum number of iteration** we should perform if convergence is taking very long to happen.It's value can be a guessed just like the value of τ in above algorithm or we can compute it's value by running the algorithm many times with different sets of initial centroid on globular data of equivalent size .Then taking mean of iterations as our guessed value for max iteration .(1)

2. **Answer 2:**
   Lines 12-16 of the algorithm describes the random initialization of the centroids. When random initialization of centroid is used, the resulting clusters are often poor. An optimum clustering can be obtained if the initial centroids lies in anywhere in a pair of clusters (since the centroids will redistribute themselves,one to each clusters). If there are k clusters then the probabaility of each cluster having exactly one centroid is much lower. Sometimes the initial centroids will readjust themselves in right way, and sometimes they dont .
   In this case,k-means will not redistribute the centroids between the pair of clusters , thus only local minimum is acheived.
   Also If we choose k centroids randomly , there are high possiblities that algorithm might choose an outlier has an initial centroid which can result into drastic increase in SSE (sum of squared errors) and also convergence might not take place.

Implications of $k-$ means

$k-$ means is simple but it is not suitable for all type of data's. Following are some drawbacks of $k-$means:

(a) Non-spherical data where mean value cannot converge towards cluster center.i.e data is non-convex.

(b) Data set having clusters of different size and density .

(c) It can not converge if the data has outliers

(d) It is only restricted to data for which can calculate centroids. (2)

3. **Answer 3:**
Time Complexity for k-means : $\Theta\ (I * K * m * n)$ where m is the number of points and n is the number of attributes. Since a bound on the iterate is included i.e. I which can be small and safely bounded,as k - means quickly reaches either complete convergence or a clustering that is close to convergence.

4. **Answer 4:**

(a) Two alternate methods for breaking *ties*:

i. Assign the datum to one of the centroid. Calculate first's cluster SSE (sum of squared errors). Remove the datum from the first cluster and place it in second cluster. Recompute the SSE for the second. The minimum of both SSE's can break the tie and datum can be placed to that cluster.

ii. Calculate each centroid cluster size. Whichever be the greater, assign the datum to that cluster.

(b) Two alternative methods for *centroidcollapses*:

i. Merging two clusters: Two clusters with closest centroid can be merged together.

ii. Re distribution of Data points : Minimum Distance parameter : In this case we will consider a *minimumdistance*. Once the centroids are updated, we will check the difference between the updated one. If the difference is less the minimum distance, we will select data points from each centroid clusters such that, selected data point is farthest one from respective centroid's. Now we will select those two data points as new centroids and again cluster the remaining point based on these newly updated one.
In this way, we are re distributing the points again making sure that number of k remains same.

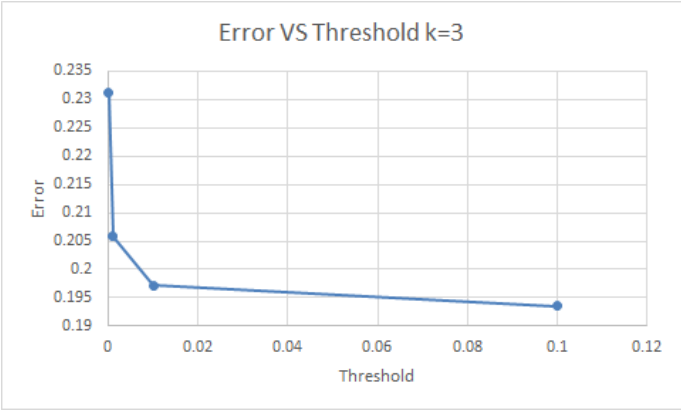

Figure 1: Fig.1 - Error Rates for different Tau

iii. τ can be determined by multiple runs on dataset given that we have fixed the number of clusters. For different τ, we can calculate error rate for fixed clusters.
The τ for which you get the least error rate can be taken.
For example: From data analysis question of this assignment, where I have implemented k- means on given breast cancer data set- I have fixed value k=3 and ran k-means for about 15 runs for different τ. Average Error rate is calculated for each τ. In figure above, It can be inferred that after τ_c (point after the elbow in the figure) , error rate is almost constant. Hence that value of τ after which there's no much change in error rate can be taken as our τ_c

(c) **pseudo-code for k-meansr - Modified the k- means to address ties and collapsing centroids**

ALGORITHM `k-means`
INPUT (data Δ, distance $d : \Delta^2 \to \mathbb{R}_{\geq 0}$, centoid number k, threshold τ)
OUTPUT (Set of centoids $\{c_1, c_2, \ldots, c_k\}$)

*** $Dom(\Delta)$ denotes domain of data.

*** Assume centroid is structure $c = (v \in DOM(\Delta), B \subseteq \Delta)$
*** $c.v$ is the centroid value and $c.B$ is the set of nearest points.
*** c^i means centroid at i^{th} iteration.

$i = 0$
*** Initialize Centroids
for $j = 1, k$ **do**
    $c_j^i.v \leftarrow random(Dom(\Delta))$
    $c_j^i.B \leftarrow \emptyset$
end for
for $j = 1, k-1$ **do** ▷ Checking what if the distance between any two centroid is < minimum distance
    for $l = j+1, k$ **do**
        if d($c_j^i.v, c_l^i.v$)<X **then** ▷ Re initialize $c_l^i.v$
            $c_l^i.v \leftarrow random(Dom(\Delta))$
        end if
    end for
end for

$f_i = \Sigma_{j=1}^{k} \Sigma_{\ell=1}^{k} d(c_j^i.v, random(Dom(\Delta))$ ▷ Computing the difference between past centroids and current
Repeat

repeat
    $i \leftarrow i + 1$
    *** Assign data point to *nearest* centroid
    for $\delta \in \Delta$ **do**
        if $(d(\delta, c_i) = d(\delta, c_j))$ **then** ▷ When ties occur
            {
            *** assignδ to firstc_i
            $SSE_1 = \sum_{z=1}^{k}((x_z - c_i)^2)$
            **** remove δ from firstc_i and assign it to δ to c_j
            $SSE_2 = \sum_{y=1}^{k}((x_y - c_j)^2)$
            $c_j^i.B \leftarrow c.B \cup \{\delta\}$, where min (SSE_1, SSE_2)
            } ▷ For handling *ties*, we will assignδ to that centroid which gives us

minimum SSE
else
$c_j^i.B \leftarrow c.B \cup \{\delta\}$, where $\min_{c_j^i}\{d(\delta, c_j^i.v)\}$
end if
end for
for $j = 1, k$ **do**
*** Get size of centroid
$n \leftarrow |c_j^i.B|$
*** Update centroid with average
$c_j^i.v \leftarrow (1/n)\sum_{\delta \in c_j^i.B} \delta$
*** Remove data from centroid
$c_j^i.B \leftarrow \emptyset$
end for
for $j = 1, k-1$ **do** ▷ To check if the distance between two centroid is < minimum distance
for $l = j+1, k$ **do**
if d($c_j^i.v, c_l^i.v$)<X **then** ▷ find farthest point from centroid $c_j^i.v, c_l^i.v$ within cluster
while true **do** ▷ j and l respectively. let the points are $p_j^i.Bandp_l^i.B$
$p_j^i.v \leftarrow \text{Max}\{d(c_j^i.B, c_j^i.v)\}$
$p_l^i.v \leftarrow \text{Max}\{d(c_l^i.B, c_l^i.v)\}$
if d($c_j^i.v, p_j^i.B$)+ d($c_j^i.v, p_j^i.B$)$\leqslant$ d($p_j^i.v, p_l^i.v$) **then**
break
end if
end while
end if
end for
end for

*** Calculate scalar product (abuse notation and structure slightly)
*** See notes
until $((1/k)\sum_{j=1}^{k} ||c_j^{i-1} - c_j^i||) < \tau$
return $(\{c_1^i, c_2^i, \ldots, c_k^i\})$

Integration
New Metric

1. Let a,b are data points
2. a_x, b_x represents their X attributes
3. a_y, b_y represents their Y attributes
4. $d_x(a_x, b_x)$ is difference between x attributes
5. $d_y(a_y, b_y)$ is difference between y attributes

Let's define a metric :
$d_x(a_x, b_x) = |a_x - b_x|$

$$d_y(a_y, b_y) = \begin{cases} 0, & (a_y = b_y) \\ 1, & (a_y \neq b_y) \end{cases}$$

Hence, our metric for this system:
$d(a, b)$=$|a_x - b_x|$+$d_y(a_y, b_y)$
Let us prove that it's a distance metric:

(a) identity of indiscernibles :
$d(a,a) = |a_x - a_x| + d_y(a_y, a_y) = 0$

(b) Symmetry :
$d(a,b) = |a_x - b_x| + 1$
$d(b,a) = |b_x - a_x| + 1$
$d(a,b) = d(b,a)$

(c) Triangular Inequality:
$d(a,b) = |a_x - b_x| + 1$
$d(b,c) = |b_x - c_x| + 1$
$d(a,c) = |a_x - c_x| + 1$
$d(a,b) + d(b,c) = |a_x - b_x| + 1 + |b_x - c_x| + 1\ a_x - c_x + 2 \geq |a_x - c_x| + 1$
Therefore,$d(a,b) + d(b,c) \geq d(a,c)$

Given the following-

$$d_1(x,y) = \begin{cases} 0, & x = y \\ 1, & x \neq y \end{cases} \quad \text{For objects } x, y.$$

$$J(x,y) = |x \cap y| / |x \cup y| \quad \text{For sets } x, y.$$
$$d_2(x,y) = 1 - J(x,y) \quad \text{For sets } x, y.$$

$$c(x,y) = \begin{cases} 0, & x = y \\ 1, & otherwise \end{cases} \quad \text{for individual characters, } e.g., \texttt{a} = \texttt{b}$$
$$d_3(\mathbf{x},\mathbf{y}) = \Sigma_{i=0}^{n-1} c(\mathbf{x}[i], \mathbf{y}[i]) \quad n = ||\mathbf{x}||, \text{the length of the string.}$$

$$d_4(\mathbf{x},\mathbf{y}) = \left| \frac{\mathbf{x}^\mathsf{T}\mathbf{y}}{||\mathbf{x}||\,||\mathbf{y}||} \right| \quad \text{for vectors } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$$

1. (a) i. $d_1(x,w)$=1

ii. $d_2(x,w)$=1-$J(x,w)$
$J(x,w) = |x \cap w| / |x \cup w|$
$|x \cap w|$=2
$|x \cup w|$=6
$J(x,w) = \frac{2}{6} = \frac{1}{3}$
$d_2(x,w)$=1-$\frac{1}{3}$
$= \frac{2}{3}$

iii. $d_3(x,w) = c(a,a) + c(b,d) + c(c,f) + c(d,e)$
$=0+1+1+1$
$= 3$

iv. $d_4(x,w) = \left| \frac{(a^2+bd+cf+de)}{\sqrt{((b^2+a^2+c^2+d^2)(a^2+d^2+f^2+e^2))}} \right|$
a=1,b=2,c=3,d=4,e=5,f=6
$= \frac{47}{2340} = 0.02005$

(b) For x = {a,b,c,d}, z = {b,f}

i. $d_1(x,z)$=1

ii. $d_2(x,z)$=1-$J(x,w)$
$J(x,z) = |x \cap z| / |x \cup z|$
$|x \cap z|$=1
$|x \cup z|$=5
$J(x,z) = \frac{1}{5}$

$d_2(x,z)$=1-$\frac{1}{5}$
=$\frac{4}{5}$

iii. $d_3(x,z)$: Since x has 4 values where as z has 2 values. We need to equal the dimentionality of x and z first. Now to calculate the missing attributes of z , we can calculate the most frequently occuring value for that attribute and replace it with that value
if the mode of one attribute is 0, replace with a global constant , "0"
In this case, $d_3(x,z)= c(a,b) + c(b,f)+c(c,0)+c(d,0) = 4$.
Another way of dealing with missing values is that we can replace missing values with the most probable value.

iv. $d_4(x,z)$=
After replacing the $w = (b,f,0,0)$,x^T w will be $ab+bf$
$\|\mathbf{x}\|$ will be $\sqrt{a^2+b^2+c^2+d^2}$
$\|\mathbf{w}\|$ will be $\sqrt{b^2+f^2}$
d_4 value will always lies between [0,1]. (explained in next question)
For example :Giving the numerical value for alphabets
a=1,b=2,c=3,d=4,f=6
$d_4(x,z)$=0.0116

In this case, d_4 is the minimum among all.

(c) Maximum value for d_1 can be either 0 or 1. Similary, for d_2 , the value of distance will lie between [0,1]. d_2 will never be greater than 1.
In calculating d_4 , denominator is a muplication of two vector norm and numerator is a inner product of two vectors. Denominator will be either greater or equal to (equal in case if both vectors are equal) than numerator.
There by, d_4 for any two sets will be greater than 0 but less than or equal to 1. It can't be greater then 1. But for d_3 for any sets- there are two possibilities :

i. None of the attributes of both of the sets are equal - $d_4 = \|\mathbf{x}\|$: length of the first set.
ii. Both the sets have same values in the attributes - $d_4 = 0$ Therefore we can say, distance d_4 lies between $0 \leq d_4 \leq$ n
Therefore, d_3 will gives us the maximum distance for any pairs only if both pair values are unequal.

(d) **F**alse
For any set v , let's calculate all the distances

i. $d_1 = 0$ since both objects are same.
ii. d_2 = 1 - 1 since,$J(v,v)$=1
d_2= 0
iii. $d_3 = 0$ since for invidual characters $c(v,v)$=0
iv. d_4:

$$d_4(\mathbf{v},\mathbf{v}) = \left| \frac{\mathbf{v}^\mathsf{T}\mathbf{v}}{||\mathbf{v}||\,||\mathbf{v}||} \right|$$

for example: let's take set x={b,f}
x^T x will be $b^2 + f^2$
$\|\mathbf{x}\|$ will be $\sqrt{b^2+f^2}$
d_4=$\frac{b^2+f^2}{\sqrt{((b^2+f^2)(b^2+f^2))}} = 1$

Therefore d_1=d_2=$d_3 \neq d_4$

2. Why combining metrics is important to integration?
Answer "Data integration is the process of integrating data from multiple sources and probably have a single view over all these sources and answering queries using the combined information" (3)
The main problem one can face in data integration is that all the data will not be of same type. This issue is known as *heterogenity* problem. There are three types of *heterogenity* problems:

(a) Data Type Heterogeneity :Storing same data with different data types

(b) Value Heterogeneity : Same logical values stored in different ways

(c) Semantic Heterogeneity : Same values in different sources can mean different things

We can combine metrics so that we can overcome heterogeneity problems in real life data set. Upon integrating, the data can be transformed accordingly.

(a) $d_{i'}(x,y) = \frac{d_i(x,y)}{1+d_i(x,y)}$ for every i.

i. identity of indiscernibles $d(x,x)$=0
Since $d_i(x,x)$ is a metric. $d_i(x,x)$ =0
$d_{i'}(x,x) = \frac{d_i(x,x)}{1+d_i(x,x)} = 0$
$d_{i'}(x,x)$=0

ii. Symmetry: $d(x,y)$=$d(y,x)$
Since $d_i(x,y)$ is a metric. $d_i(x,y) = d_i(y,x)$
Therefore:
$d_{i'}(x,y) = \frac{d_i(x,y)}{1+d_i(x,y)}$
$=\frac{d_i(y,x)}{1+d_i(y,x)} = d_{i'}(y,x)$
$d_{i'}(x,y) = d_{i'}(y,x)$

iii. Triangle Inequality: $d(x,y) + d(y,z) \geq d(x,z)$
Since $d_i(x,y)$ is a metric. It will satisfy the triangle inequality property
$=d_i(x,y) + d_i(y,z) \geq d_i(x,z)$
divide each value with $1 + d_i(x,y)$
$=\frac{d_i(x,y)}{1+d_i(x,y)} + \frac{d_i(y,x)}{1+d_i(y,x)} \geq \frac{d_i(x,z)}{1+d_i(x,z)}$ viz. $d_{i'}(x,y) + d_{i'}(y,z) \geq d_{i'}(x,z)$

Hence $d_{i'}(x,y) = \frac{d_i(x,y)}{1+d_i(x,y)}$ is a distance metric
For example:
For x, y are two disjoints sets

A. $d_{1'}(x,x)$=0 since $x = x, d_1(x,x) = 0$

B. $d_1(x,y) = 1$
$d_{1'}(x,y) = \frac{d_1(x,y)}{1+d_1(x,y)} = 1/2$
$d_1(y,x) = 1$
$d_{1'}(y,x) = \frac{d_1(y,x)}{1+d_1(y,x)} = 1/2$
$d_{1'}(x,y) = d_{1'}(y,x)$

C. For x, y, z three disjoint sets. $d_{1'}(x,y) = \frac{d_1(x,y)}{1+d_1(x,y)} = 1/2$
$d_{1'}(y,z) = \frac{d_1(y,z)}{1+d_1(y,z)} = 1/2$
$d_{1'}(x,z) = \frac{d_1(x,z)}{1+d_1(x,z)} = 1/2$
$d_{1'}(x,y) + d_{1'}(y,z) \geq d_{1'}(x,z)$

(b) $d_{i'}(x,y) = \alpha d_i(x,y)$ for $\alpha \in \mathbb{R}_{>0}$

i. identity of indiscernibles $d(x,x)$=0
Since $d_i(x,x)$ is a metric. $d_i(x,x)$ =0
$d_{i'}(x,x) = \alpha d_i(x,y) = 0$
$d_{i'}(x,x)$=0

ii. Symmetry: $d(x,y)$=$d(y,x)$
Since $d_i(x,y)$ is a metric. $d_i(x,y) = d_i(y,x)$
Therefore:
$d_{i'}(x,y) = \alpha d_i(x,y)$ =$\alpha d_i(y,x) = d_{i'}(y,x)$
$d_{i'}(x,y) = d_{i'}(y,x)$

iii. Triangle Inequality: $d(x,y) + d(y,z) \geq d(x,z)$
Since $d_i(x,x)$ is a metric. It will satisfy the triangle inequality property
=$d_i(x,y) + d_i(y,z) \geq d_i(x,z)$
multiply by α to the whole equation:
=$\alpha d_i(x,y) + \alpha d_i(y,z) \geq \alpha d_i(x,z)$
viz. $d_{i'}(x,y) + d_{i'}(y,z) \geq d_{i'}(x,z)$
Hence, $d_{i'}(x,y) = \alpha d_i(x,y)$ is a distance metric

(c) $d_5(x,y) = d_1(x,y) + 3d_2(x,y)$

i. identity of indiscernibles $d(x,x)$=0
Since $d_1(x,y), d_2(x,y)$ are metrics. It has to satisfy identity of indiscernibles property.
Hence, $d_1(x,x), d_2(x,x) = 0$
Therefore, $d_5(x,x) = d_1(x,x) + 3d_2(x,x) = 0$

ii. Symmetry: $d(x,y)$=$d(y,x)$
Since $d_1(x,y) and d_2(x,y)$ are metrics. It has to satisfy symmetry property.
Hence, $d_1(x,y) = d_1(y,x) and d_2(x,y) = d_2(y,x)$
Therefore, $d_5(x,y) = d_1(x,y) + 3d_2(x,y) = d_1(y,x) + 3d_2(y,x) = d_5(y,x)$

iii. Triangle Inequality :
Since d_1 and d_2 are metric, both will satisfy Triangular inequality property
$d_5(x,y) + d_5(y,z)$=$d_1(x,y) + 3d_2(x,y) + d_1(y,z) + 3d_2(y,z)$
= $((d_1(x,y) + d_1(y,z)) + 3(d_2(x,y) + d_2(y,z))$
=$d_1(x,z) + 3d_2(x,z) = d_5(x,z)$
Hence, d_5 is a metric

(d) $d_6(x,y) = d_2(y,x)$ Since d_2 is a distance metric and it will satisfy all three properties of distance metric.

i. Identity of indiscernibles
$d_6(x,x) = d_2(x,x) = 0$

ii. Symmetry:
$d_6(x,y) = d_2(y,x) = d_2(x,y) = d_6(y,x)$
$d_6(x,y) = d_6(y,x)$

iii. Triangle Inequality :
$d_6(x,y) + d_6(y,z)$= $d_2(y,x) + d_2(z,y)$
=$d_2(x,y) + d_2(y,z)$ $(since, d_2(x,y) = d_2(y,x))$
$\geq d_2(x,z) \geq d_6(x,z)$
Hence d_6 is a metric

(e) $d_7(x,y) = d_3(x,y)d_2(x,y)$

i. Identity of indiscernibles
$d_7(x,x) = d_3(x,x)d_2(x,x) = 0$

ii. Symmetry:
Since$d_3(x,y) = d_3(y,x)$ Mulitply with $d_2(x,y)$ with whole equation:
$d_3(x,y).d_2(x,y) = d_3(y,x).d_2(x,y)$
$d_3(x,y).d_2(x,y) = d_3(y,x).d_2(y,x)$ (since, $d_2(x,y) = d_2(y,x)$)
=$d_7(x,y) = d_7(y,x)$

iii. Triangle Inequality :
$d_7(x,y) + d_7(y,z) = d_3(x,y)d_2(x,y) + d_3(y,z)d_2(y,z)$
$\geq d_7(x,z)$ Hence d_7 is a metric

3. "A survey tree edit distance and related problems"
The whole document explains on the importance of tree labeling and local operations of nodes labeled among them to perform basic operations of deletion, insertion and relabel to simplify solving problems. These help us relate various problems to tree edit distance, alignment distance and inclusion. The key concepts of Tree edit distance and how it is calculated are discussed. With a cost function defined, The $TreeEditDistance$ is the minimum cost of the optimal edit script between two labeled trees. The $TreeAlignmentDistance$, with a cost function defined on labels, is the minimum cost of the optimal alignment between two labeled trees. Two trees can be called Inclusive if one of the trees can be obtained by deleting nodes of the second tree. The rest of the document summarizes on how the various Edit distances are derived from the various algorithms mentioned. These Edit distances, Alignment distances and Inclusions come with different variants as mentioned in the table below **Tree Distance Figure**:
Legend: Results for the tree edit distance, alignment distance, and inclusion problem listed

Tree edit distance

variant	type	time	space
general	O	$O(\|T_1\|\|T_2\|D_1^2D_2^2)$	$O(\|T_1\|\|T_2\|D_1^2D_2^2)$
general	O	$O(\|T_1\|\|T_2\|\min(L_1,D_1)\min(L_2,D_2))$	$O(\|T_1\|\|T_2\|)$
general	O	$O(\|T_1\|^2\|T_2\|\log\|T_2\|)$	$O(\|T_1\|\|T_2\|)$
general	O	$O(\|T_1\|\|T_2\| + L_1^2\|T_2\| + L_1^{2.5}L_2)$	$O((\|T_1\| + L_1^2)\min(L_2,D_2) + \|T_2\|)$
general	U	MAX SNP-hard	
constrained	O	$O(\|T_1\|\|T_2\|)$	$O(\|T_1\|\|T_2\|)$
constrained	O	$O(\|T_1\|\|T_2\|I_1I_2)$	$O(\|T_1\|\|D_2I_2)$
constrained	U	$O(\|T_1\|\|T_2\|(I_1+I_2)\log(I_1+I_2))$	$O(\|T_1\|\|T_2\|)$
less-constrained	O	$O(\|T_1\|\|T_2\|I_1^3I_2^3(I_1+I_2))$	$O(\|T_1\|\|T_2\|I_1^3I_2^3(I_1+I_2))$
less-constrained	U	MAX SNP-hard	
unit-cost	O	$O(u^2\min(\|T_1\|,\|T_2\|)\min(L_1,L_2))$	$O(\|T_1\|\|T_2\|)$
1-degree	O	$O(\|T_1\|\|T_2\|)$	$O(\|T_1\|\|T_2\|)$

Tree alignment distance

general	O	$O(\|T_1\|\|T_2\|(I_1+I_2)^2)$	$O(\|T_1\|\|T_2\|(I_1+I_2)^2)$
general	U	MAX SNP-hard	
similar	O	$O((\|T_1\|+\|T_2\|)\log(\|T_1\|+\|T_2\|)(I_1+I_2)^4s^2)$	$O((\|T_1\|+\|T_2\|)\log(\|T_1\|+\|T_2\|)(I_1+I_2)^4s^2)$

Tree inclusion

general	O	$O(\|T_1\|\|T_2\|)$	$O(\|T_1\|\min(D_2L_2))$
general	O	$O(\|\Sigma_{T_1}\|\|T_2\| + m_{T_1,T_2}D_2)$	$O(\|\Sigma_{T_1}\|\|T_2\| + m_{T_1,T_2})$
general	O	$O(L_1\|T_2\|)$	$O(L_1\min(D_2L_2))$
general	U	NP-hard	

Figure 2: Tree Distance

according to variant. D_i, L_i, and I_i denotes the depth, the number of leaves, and the maximum degree respectively of T_i, i = 1, 2. The type is either O for ordered or U for unordered. The value u is the unit cost edit distance between T_1 and T_2 and the value s is the number of spaces in the optimal alignment of T_1 and T_2.
The type is either O for ordered or U for unordered. The value u is the unit cost edit distance between T_1 and T_2 and the value s is the number of spaces in the optimal alignment of T_1 and T_2.
The value of $\sum T_1$ is set of labels used in and T_2. is set of labels used in T_1 and $m_(T_1, T_2)$ is the number of pairs of nodes in T_1 and T_2 which have the same label.

Applications of k-means and Data Preparation to Medical Data

1. **Data Mining Problem**

(a) Average cost of biopsy: (1000+5000)/2 = 3000
There are total 699 patients on which biopsy has to be done.
Total cost for biopsy will be 699*3000 = 2097000 dollars.
The cost of computer program will be 699*10= 6990 dollars

(b) Average cost of masectomy = (15000+55000)/2= 35000 dollars. Number of malignant cases among the data set = 241
Total cost for a malignant patient : 241 * 35000 = 8435000 dollars

(c) There are total 241 patients who have malignant tumor. Since our computer program efficiency is about 90%, There are high chances that 10% of these 241 patients might be clustered with benign class i.e. 24 patients will be left without any treatment.
Since the mortality rate given is 70% , total 17 people can die , if they are not treated in next five years

(d) Breast cancer is a rising issue among women. A cancer's stage is a crucial factor in deciding what treatment options to recommend, and in determining the patient's prognosis. The problem with our computer program is that it's 90% efficient.There are high chances that a women having cancer may be diagnosed as "Benign".That 10% of the patient might left untreated which is life threatening. The problem with the given data set is that total 16 patient have missing Bare Nuceli values. Every attribute is equally important for diagnosis of a Patient. Therefore we need to predict the missing values accurately and replace those with the most probable value for the given data set.
Also we have duplicate data in the data set which can lead to inaccurate data mining conclusions.

2. **Data Preparation**

(a) Removing SCN and C columns, There are total 9 attributes in data $\triangle$

(b) There are 16 missing vaues in data $\triangle$, therefore the size of $\triangle^m$is 16

(c) There are total 16 patients for whom Bare Nuclei data is missing.

(d) SCN for those who have missing values : 1057013 1096800, 1183246, 1184840, 1193683, 1197510, 1241232, 169356, 432809, 563649, 606140, 61634, 704168, 733639, 1238464, 1057067

(e) Out of these 16 women, 2 women have been detected with malignant cancer.Bare nuclei is missing for these women and also from question 3(d) where we are finding out the linearity of each attribute,we cannot remove bare nuclei before proceeding for clustering (Bare nuclei is not much linear with any other attribute such that I can remove it from the data set).Therefore, I would recommend re-examination for the women because there are high chances that a person having malignant cancer being clustered with benign group.
If all the 16 women to go re-examination, biopsies has to be performed again which will cost on an average $3000 .
so total cost of re-examination of 16 women would be 16×3000=$48000 .

Total cost of Computer program will 16×10=$160

(f) Out of 699 records only 16 records are having missing data which is around 2.29% of total available records , which is not much . Also only missing value is for attribute A_6 .i.e Bare Nuclei , which can easily be predicted using Frequency Distribution or Conditional Probablity .
Thus Amount of missing data is not significant .

(g) The tuples with unknown data should not be removed from the data set. The issue of missing data must be addressed because if we ignore this problem in data preparation stage , it can lead to introduce bias into the models laters which can lead to inaccurate data

mining conclusions. As per patient's perspective, if we remove unknown data-it can lead to two things- re-examination of the patient (which cost higher) and also misdiagnosis of the patient which is life threatening.

(h) Unknown Data can be replaced by two methods:

i. Replacing the missing values by the most frequent value:
Since A_7 has the missing values in the given data set, let us calculate the most frequently occuring value in the A_7 attribute. out of 699 record,
we have $A_6 = 1 = 402, 2 = 30, 3 = 28, 4 = 19, 5 = 30, 6 = 4, 7 = 8, 8 = 21, 9 = 9, 10 = 132$
We can see that 1 has appeared most frequently in A_6 attribute , we can assign value 1 to the missing attributes
(I have used this method to replace the missing values - DeltaFix.csv)

SCN	A_i	**data**
1238464	A_7	1
1183246	A_7	1
1184840	A_7	1
1241232	A_7	1
733639	A_7	1
563649	A_7	1
1096800	A_7	1
432809	A_7	1
61634	A_7	1
1057067	A_7	1
1057013	A_7	1
704168	A_7	1
169356	A_7	1
606140	A_7	1
1193683	A_7	1
1197510	A_7	1

ii. Other way of predicting Unknown value using Naive Bayes Classifier: For calculating the most probable value for A_7, we need to consider the following:

A. Only Attributes A_2 to A_10 is considered. SCN will not be considered for calculating conditional probability.

B. All the attributes are independent of each other.

Having above two assumptions, we can predict the missing value of A_7 by using the following formula:

$$P(A_7 \mid X) = \frac{P(X|A_7) \times P(A_7)}{P(X)}$$

Where,
$P(A_7 \mid X)$ = posterior probability.
$P(A_7)$ = Prior probability.
$P(X \mid A_7)$= Class conditional probability.
$P(X)$ = Evidence.

Although this method is more complex, but can predict the missing values more accurately.

3. **Data Analysis**

(a) R code - DataAnalysisCode.R - connection between SQL and R

(b) R code - PlotForEachAttribute.R -
Plots for each attribute in fig 3 below

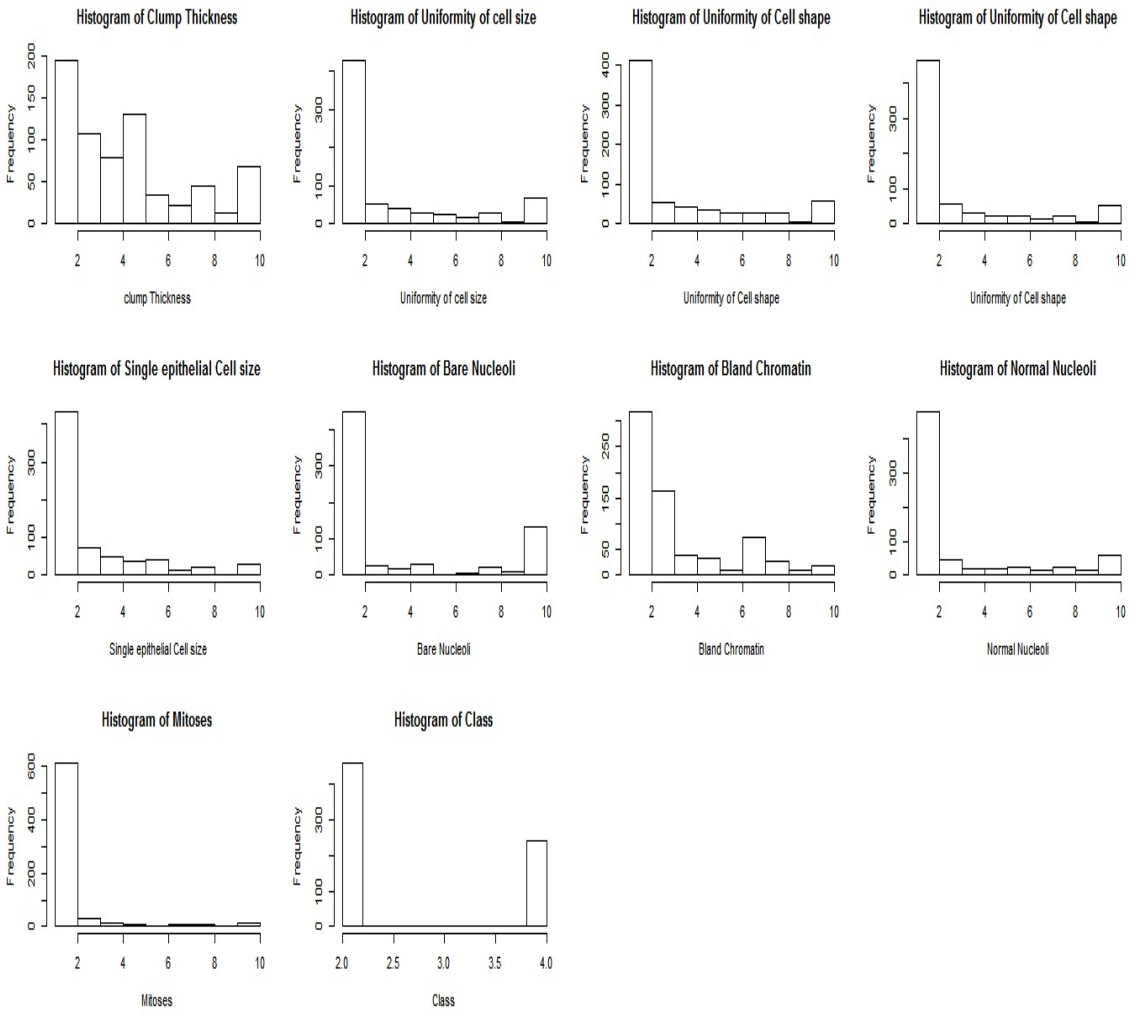

Figure 3: Plots for Each Attributes

Please find "plots for each attributes.jpg" in HW folder if you find the pic in pdf is not clear .

(c) R Code- DeltaAnalysisCode- Mean , Median , Mode and Variance for each attribute: Fig4

Attributes	Mean	Median	Mode	Variance
A2	4.41774	4	1	7.9284
A3	3.134478	1	1	9.3114
A4	3.207439	1	1	8.83227
A5	2.806867	1	1	8.15319
A6	3.216023	2	2	4.90312
A7	3.486409	1	1	13.1139
A8	3.437768	3	2	5.94562
A9	2.866953	1	1	9.32468
A10	1.589413	1	1	2.94149

Figure 4: Mean,Median,Mode and Variance

(d) **Pearson Correlation Coefficient :**

Pearson Correlation Coefficient gives a linear dependence between two Variable X and Y whose value lies between +1 and -1, where +1 gives total positive correlation (all the point lies on a line for which Y increases X increases), 0 gives no correlation and -1 gives total negative correlation.
$cos\theta = \frac{\mathbf{xy}}{||\mathbf{x}||||\mathbf{y}||}$ i.e. cosine similarity is nothing but the similarity between two non zero vectors that gives us cosine of angle between them.
If the angle between two vectors is 0 then two vectors have same orientation. If the angle between two vectors is 90 then two vectors have 0 similarity. Therefore, Pearson Correlation can be also viewed as as the cosine of angle θ between two vectors of sample in N-dimension.

Purpose of this step:

The main purpose of this step is to reducing the dimentionality of given data set. The dimensionality reduction will remove irrelevant , weakly relevant and redundant attributes.
R-Code- DeltaClean.R

```
"A:2:A:3:0.64491250435127"
"A:2:A:4:0.654589080001924"
"A:2:A:5:0.486356243676702"
"A:2:A:6:0.521816219959853"
"A:2:A:7:0.590008211356896"
"A:2:A:8:0.558428162285396"
"A:2:A:9:0.535834549212977"
"A:3:A:4:0.906881913052594"
"A:3:A:5:0.705581811557112"
"A:3:A:6:0.751799129877131"
"A:3:A:7:0.686672799928519"
"A:3:A:8:0.755720981100574"
"A:3:A:9:0.722864821906358"
"A:4:A:5:0.683079200230476"
"A:4:A:6:0.719668437170359"
"A:4:A:7:0.70747376648218"
"A:4:A:8:0.735948454023297"
"A:4:A:9:0.719446316953281"
"A:5:A:6:0.599599068425499"
"A:5:A:7:0.666971029642029"
"A:5:A:8:0.666715326264053"
"A:5:A:9:0.603352412216761"
"A:6:A:7:0.58370144812926"
"A:6:A:8:0.616101840871849"
"A:6:A:9:0.628880685589092"
"A:7:A:8:0.674214707285972"
"A:7:A:9:0.574778133132502"
"A:8:A:9:0.665877809425439"
```

Figure 5: Pearson Coefficient

4. I have implemented $k-$ means using java. I have attached $k-$Means Java file in my homework folder.
   There are total 7 java files inside. One has to import my java folder into IDE and check my code.
   Plots for k- means:
   As we can see from our plot between error rate and k for constant τ, the error rate is increasing linearing with the increase in number of clusters. It can be inferred from the graph that as the number of clusters is increasing, the malignant data is clustered with benign data which is resulting to higher error rate. In that case, a patient with malignant tumor will be placed in benign cluster. This will lead to misdiagnosis of the patient which is life threatening. Average value of error rate for 20 runs is given Fig-7

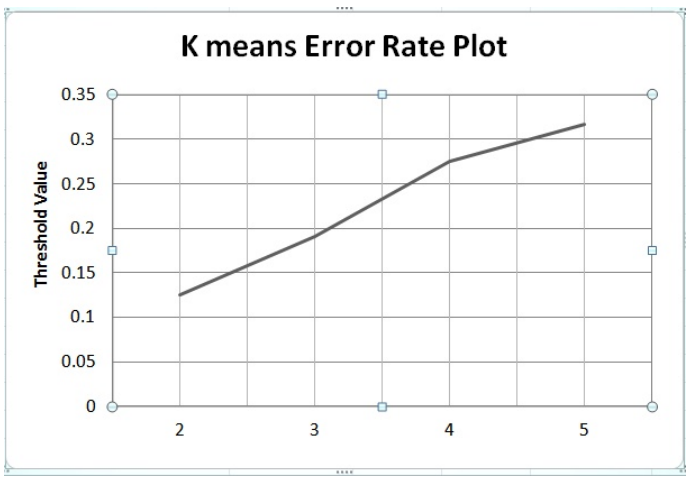

Figure 6: K-Means v/s error rate

tc=0.001	k=3	k=4	k=5	k=2
Multiple runs for k=2,3,4,5 for same threshold value	0.255885	0.28198	0.27265	0.13360524
	0.1586452	0.25574	0.27952	0.1354062
	0.1586452	0.27581	0.28296	0.1354062
	0.24988127	0.24705	0.37674	0.13444887
	0.15864521	0.35318	0.25894	0.1354062
	0.2378129	0.34903	0.18188	0.135406
	0.15877837	0.1656	0.25809	0.13168539
	0.15877837	0.27198	0.36772	0.1354062
	0.15864521	0.27337	0.3805	0.12088336
	0.1586452	0.27978	0.29543	0.1354062
	0.1354062	0.24705	0.36772	0.07528789
	0.23570561	0.27265	0.37674	0.10070081
	0.1586452	0.23571	0.36406	0.1345429
	0.2498812	0.34903	0.36406	0.1354062
Average	**0.19041192**	**0.27557**	**0.31622**	**0.12568211**

Figure 7: Average error rate for all the k=2,3..5

Reference:
1,2-ISBN:9780321321367 Introduction to Data mining
3-web.cs.wpi.edu/ cs561/s12/Lectures/IntegrationOLAP/DataIntegration.pdf